



# Aihemallinnuksen mahdollisuudet sosiaalisen median aineistojen jäsentämisessä – terveyskeskustelu Suomi24-verkkopalstalla

Marjoriikka Ylisiurua

## TIIVISTELMÄ

Aihemallinnus on koneoppimisen menetelmä, joka mahdollistaa suurten tekstimassojen automatisoidun tarkastelun. Artikkelissani kokeilen suosittua LDA-aihemallinnusmenetelmää ja kartoitan mahdollisuuksia tiivistää suuria aineistoja tulkinallisen tutkimuksen tarpeisiin. Aineisto koostuu suomenkielisen Suomi24.fi-keskustelupalstan viesteistä. Aihemallinnus löytää tekstistä paitsi tiedossa olevia ja tuntemattomia keskustelunaiheita, myös keskustelussa yleisesti toistuvia fraaseja. Täydentämällä tutkimusprosessia lähilukemalla on lisäksi mahdollista astua aiheiden ulkopuolelle ja löytää tarinallisesti etenevä, keskustelun yleinen narratiivi. Lähiluvun laaja hyödyntäminen liudentaa aihemallinnuksen määrällistä ja jopa jäykkää leimaa ja tarjoaa sen "neutraalien" ratkaisujen tuottamisen sijaan tulkitsevaan tutkimukseen sopivaksi aineiston tiivistämisen menetelmäksi.

## Johdanto

Internetissä tapahtuva sosiaalinen vuorovaikutus on kasvanut tärkeäksi osaksi ihmisten elämää. Kirjoittelun, jakamisen ja peukutusten seurauksena ystäväpiirien virtuaalisiin hampaisiin pääsee laaja kirjo elämän osa-alueita, kuten sote-uudistuksen poliittiset vaiheet tai itse kunkin arjen vastoinikäymiset ja ilot. Tutkijoiden kannalta Internetissä syntyneet aineistot mahdollistavat moninaisten uusien kysymysten esittämisen. Yksi verkkoaineistojen ja etenkin sosiaalisen median keskusteluaineistojen eduista on niiden antama näkökulma ihmisten keskinäiseen vuorovaikutukseen ja sen suhteeseen ympäröivän maailman tapahtumiin. Verkkoaineistoja hyödynnetäänkin nykyään niin Internet-ilmiön tutkimisessa ja vanhojen teorioiden todentamisessa kuin yksinkertaisten luonnontieteen menetelmillä mitattujen ilmiöiden, kuten verkostorakenteiden ja sentimenttien, leviämisessä (Gil De Zúñiga ja Diehl 2017).

Monet verkkosisällöt koostuvat kirjoitetusta tekstistä, ja useat tutkijat ovatkin

lähestyneet verkossa ilmeneviä ilmiöitä lähilukemalla sosiaalisen median sisältöjä (Dolbec ja Fischer 2015; Huovila ja Saikkonen 2016; Jauho 2016; Parmentier ja Fischer 2014; Veer, Ozanne ja Hall 2016). Aineistojen laajuudesta johtuen niitä on kuitenkin usein taloudellista analysoida automatisoidusti tietokoneohjelmistoilla (DiMaggio, Nag ja Blei 2013; Guggenheim ym. 2015). Tällaisten menetelmien toimivuutta on arvioitu myös Suomessa (Ahonen 2015; Haverinen ja Suominen 2015; Parviainen ym. 2012; Purhonen ja Toikka 2016; Tolonen ja Lahti 2015; Winter ja Wiberg 2016).

Tutkin artikkelissani automatisoidun analyysin tarjoamia mahdollisuuksia verkkoaineistojen tulkinnallisessa tutkimuksessa. Tulkinnallisuuden yhdistäminen ”koneen oppimiseen” saattaa tuntua ristiriitaiselta, sillä koneoppimisen keinoin toteutettu automatisoitu etälukeminen eroaa ihmistieteissä tavanomaisesta lukemisen tavasta, ns. lähilukemisesta (Berry 2011). Tekstin muuntaminen taulukkomuotoon hävittää mahdollisuuden tulkita esimerkiksi aikakauslehden tekstiä suhteessa viereisen sivun kuvaan. Monet vuorovaikutuksen vivahteet myös jäävät koneoppimismenetelmillä huomioimatta. Kun tutkimuksen kohteena ei ole tekstidokumentti tai siinä käytetyt sanat vaan vuorovaikutus, ihminen on edelleen ylivoimainen löytämään ja tekemään tulkintoja (Savage 2013; Su ym. 2017).

Tämän artikkelin kirjoitushetkellä vaikuttaa siltä, että ainakin yhdestä asiasta on päästy yksimielisyyteen: samalla tavoin kuin ”perinteinen” laadullinen ja määrällinen tutkimus ovat eläneet rinta rinnan, automatisoitu etäluku ei korvaa lähilukua. Koneoppiminen ei kuitenkaan ole täysin lähiluvun tulkintaperinteistä erillinen saati yhtenäinen menetelmäkokonaisuus, vaan sen piirissä on esimerkiksi tulkinnallisuutta enemmän korostavia menetelmiä, ja analyysin tulokset voidaan itsessään hahmottaa tulkinnallisuuksiksi (Ahonen 2015; Buurma 2015). Lähilukuun tottunut tutkija voi havaita, että automatisoitu etäluku löytää uudennlaisista aineistoista tarkkuutta ja erityisyyttä, ei vain yleistyksiä, ja voi jopa soveltaa sekä etä- että lähilukua samassa tutkimuksessa.

Varsin vakiintunut menetelmä laajojen aineistojen automatisoituun analysointiin on LDA-aihemallinnusmenetelmä (Latent Dirichlet Allocation). Aihemallinnus systematisoi aineiston merkityksiä induktiivisesti havainnoista tunnistettujen aiheiden perusteella (Purhonen ja Toikka 2016). LDA tarkastelee tekstin merkitysten suhteellisuuksia sen sanojen suhteiden kautta sekä havaitsee diskursiivisia konteksteja ja niiden ajallisia muutoksia (DiMaggio ym. 2013: 578, 586–587, 593–594). Yhtenä LDA-mallinnuksen rajoituksena on sen pyrkimys luokitella tekstistä varsin rajattuja yksiköitä. Mallinnuksen aikana lauseet rikotaan osiinsa eli sanoihin. Näin esimerkiksi sanajärjestys, sanamuodot tai yksittäisen sanan lähikonteksti eivät vaikuta aihemallinnuksen tuloksiin.

LDA-menetelmän rajoituksista huolimatta sitä hyödyntävä tulkinnallista analyysiä tekevä tutkija voi paikallistaa keskustelunaiheita massiivisista tekstikokoelmista, joiden perinteinen lähiluku ja koodaus veisivät aineistojen koon vuoksi huomattavasti aikaa. Teknisesti LDA on suuntaamaton koneoppimismenetelmä

(Blei 2012; Chang ym. 2009), joka muistuttaa tuttua tekstin klusterianalyysia (vrt. Heinonen ym. 1996). Tämä tarkoittaa sitä, ettei tutkijalla ole etukäteen mahdollisuuksia vaikuttaa siihen, mitä mallinnusohjelma etsii, esimerkiksi tarjoamalla sille valikoituja avainsanoja. Tutkijalla on mahdollisuus valita etsittävien aiheiden määrä. Tällöin tutkijan on analysoitava aineisto useampaan kertaan valitsemalla eri lukumäärä etsittäviä aiheita. Tutkijan tulkintaa on paitsi se, kuinka monta perusteltua aihetta menetelmä onnistui löytämään, myös niiden kuvailu sekä aiheiden nimeäminen.

Oma tutkimukseni hyödyntää LDA-menetelmää aineiston tiivistämisessä, keinona paikallistaa kiinnostavia verkkovuorovaikutuksen osia. Automatisoidun analyysin jälkeen ja rinnan sen kanssa on kuitenkin tehtävä eräänlainen laadullinen loikka takaisin kohti alkuperäistä tekstiä kontekstissaan. Laadullinen loikka murtaa tyypillisen tavan hyödyntää LDA-menetelmää pelkkänä tekstien luokittelijana ja palauttaa alkuperäistekstin tulkinnan mahdollisuuksineen tutkijan ulottuville.

## Aineisto

Tutkimukseni aineiston syntypaikka on Suomi24.fi-sivusto. Se on paitsi Suomen suurin, myös yksi maamme vanhimmista Internet-keskustelupalstoista. Keskustelupalstan sisältö vuodesta 2001 nykypäivään<sup>1</sup> on rajoittamattomasti tarjolla akateemiseen käyttöön. Kaikkiaan Suomi24-aineistoa luonnehtivat seuraavat tekijät (Lagus ym. 2016):

- Keskustelupalstalla käydään aiheasuuntautunutta keskustelua.
- Kirjoittajista noin 10 % on kirjoitushetkellä rekisteröitynyt ja sisäänkirjautunut.
- Aineisto ei sisällä käyttäjiä koskevaa tietoa, kuten ikää tai sukupuolta.
- Aineisto kattaa pitkän aikavälin, vuodet 2001–2015, joka mahdollistaa 2000-luvun alun suomalaisen keskustelukulttuurin ja sen ilmapiirin tarkastelun.
- Aineiston kirjoitushetki on tallennettu sekunnin tarkkuudella, jolloin viestinnän muotoa ja sen intensiteettiä voi tarkastella eri vuorokauden tai vuodenaikoina.

Suomi24-keskustelupalsta on ainutlaatuinen aineisto, sillä tuoreetkin laajoja keskustelullisia aineistoja käsittelevät analyysit koskevat yleensä rakenteista, rajatun mittaista dataa, kuten 140 merkkiä sisältäviä Twitter-viestejä. Pienemmät algoritmisesti käsiteltävät tekstiaineistot puolestaan kattavat useimmiten poliittisia,

---

<sup>1</sup> <http://metashare.csc.fi/repository/browse/the-suomi-24-corpus/b4db73da85ce11e4912c005056be118ea699d93902fa49d69b0f4d1e692dd5f1/>

institutionaalisia tai mediasisältöjä. (Gonzalez-Bailon ja Paltoglou 2015; Hannak ym. 2012; Paul ja Dredze 2014; Schwartz ym. 2014; Weng, Menczer ja Ahn 2013; Wilkerson, Smith ja Stramp 2015)

Suomi24-keskustelupalstalta muodostettu aineisto koostuu viesteistä, joita palvelun käyttäjät kirjoittavat tai pelkästään lukevat. Viestit kirjoitetaan kävijöiden aloittamiin keskusteluketjuihin, joita ylläpitäjät tarvittaessa moderoivat. Suomi24.fi-sivuston voi nähdä yhtenä verkkoyhteisönä, mutta palstan ylläpitäjät arvioivat sen sisällä olevan lukuisia pienempiä yhteisöjä. Yhteisöjä muodostuu palstan valmiille alipalstoille, mutta keskustelijat voivat myös vaikuttaa siihen, miten alipalstoja muodostetaan, viemällä keskustelua haluamaansa suuntaan tai pyytämällä ylläpitoa perustamaan uusia palstoja (Lagus ym. 2016). Yhteisössä on yli 1 600 alipalstaa hierarkkisessa järjestyksessä (Taulukko 1), mutta näiden rajaamien aiheiden lisäksi on oletettavissa, että joistakin aiheista keskustellaan myös niille nimettyjen palstojen ulkopuolella. Sivustolla käsitellään myös aiheita, joille ei ole nimetty omaa palstaa.

Taulukko 1. Esimerkki Suomi24-keskustelupalstan hierarkiasta. Terveyspalstan alipalstojen toinen taso ja niiden kommenttimäärät.

Alipalsta	Viestien lukumäärä
Terveys > Henkinen hyvinvointi ja mielenterveys	582 428
Terveys > Sairaudet	575 150
Terveys > Painonhallinta	416 879
Terveys > Päihteen ja riippuvuudet	288 574
Terveys > Terveysten hoito	165 902
Terveys > Naisten terveys	88 535
Terveys > Yleistä terveydestä	68 802
Terveys > Ehkäisy	62 497
Terveys > Lääkkeet	60 681
Terveys > Suun terveys	60 143
Terveys > Plastiikkakirurgia	54 493
Terveys > Aistit	43 352
Terveys > Kuolema ja suru	41 563
Terveys > Kysy terveydestä	29 590
Terveys > Miesten terveys	25 741
Terveys > Kysy allergioista	1 527
Terveys > Terveystenhoitopalvelut	1 331
Terveys > Vastustuskyky	527
Terveys > Kysy plastiikkakirurgiasta	513

Menetelmän havainnollistamiseksi rajasin aineiston terveysaiheiseen keskusteluun. Lääkärin hoituhuone ei ole ainoa paikka, jossa tehdään terveyteen ja hyvinvointiin liittyviä päätöksiä. Verkkopalvelujen ja muun teknologian kehittymisen myötä kuluttajien ulottuville on tullut uusia tapoja vaikuttaa henkilökohtaiseen terveyteensä ja hyvinvointiinsa (Hardey 2001). Internetin vaikutuksia potilaskuluttajien ja lääkärin suhteisiin on aikaisemmin tutkittu (Ahmad ym. 2006;

Autio, Helovuori ja Autio 2012), mutta myös verkkovuorovaikutusta on havainnoitu koskien sosiaalisten verkostojen ja vertaistuen vaikutusta terveyteen (Centola 2013; Wright ja Bell 2003) sekä ruokavaliovalintojen ympärillä käytäviä kansalaiskeskusteluja (Huovila ja Saikkonen 2016; Jauho 2016).

Aineiston rajaamisen kannalta alipalstojen orgaanisuus on haasteellista: ilman laajempaa luotaavaa tutkimusta on vaikea luottaa siihen, että tietty ilmiö saadaan haltuun tarkastelemalla yhtä rajattua alipalstaa. Esimerkiksi painonhallintaan ja ruokavalioon liittyviä keskusteluja käydään yhtä lailla Terveys-palstan Diabetes- ja Painonhallinta-alipalstoilla kuin Ruoka ja juo-ma -palstan Ravinto ja ruokavaliot -alipalstoilla. On siis mahdollista, että jonain ajankohtana valittuun tutkimuskysymykseen olennaisesti liittyviä keskusteluja onkin käyty myös muualla kuin tutkijan mieleen ensiksi tulevalle keskustelualueella. Yksittäisten alipalstojen tutkiminen ja käsin poiminta ei siis takaa sitä, että kaikki aiheeseen liittyvät keskustelut saadaan tavoitettua.

## **Analyysin valmisteluvaiheet**

Tutkimuksessa oli seitsemän vaihetta. Ensimmäiseksi toteutin valmisteluvaiheita luettavat vaiheet: 1) luotaava koko Suomi24-aineistoon perehtyminen, 2) analyysiohjelmiston toteutus sekä 3) kohteeksi valitun terveysaineiston keruu. Vaiheita voi suorittaa eri järjestyksessä ja limittäin, mutta varsinaisessa analyysissä tarvitaan niitä kaikkia.

Seuraavissa vaiheissa 4) toteutin useita LDA-aihemallinnuksia, 5) paikallistin viestejä aiheiden avainsanojen avulla sekä luin niitä XLS-muodossa ja 6) perehdyin alkuperäiseen keskustelukontekstiin lähilukemalla Suomi24-verkkosivustoa, tarkoitukseni 7) valita lopullinen malli ja kuvailla tulokset. Valintaa edeltävät kolme vaihetta 4–6 suoritettiin limittäin, kuitenkin niin, että yhden LDA-mallin laskennan jälkeen etenin keskustelujen lähilukuun samalla kun ohjelmisto analysoi toista LDA-mallia.

Vaiheessa yksi perehdyin aineistoon Kielipankin FIN-CLARIN Korp-työkalun avulla<sup>2</sup>. Työkalulla voi toteuttaa avainsanahakuja, joiden lopputulos esitetään lauseina, joissa etsitty sana esiintyy. Yksinkertaisia sanahakutoimintoja käyttämällä tarkastelin kiinnostavan aiheen esiintyvyyttä keskustelussa, keskustelutyylillä, ajankohtaa sekä kontekstia. Sanahakua kokeilemalla oli lisäksi helpompi hahmottaa, millaisia elementtejä algoritmisella analyysillä kenties pystyisi tunnistamaan. Yksinkertainen sanahaku on nopea, mutta karkea menetelmä aineiston mahdollisten sisältöjen hahmottamiseen (Savage 2013). Aineiston ymmärtämiseksi tarkastelinkin samaan aikaan koko keskustelupalstan sisältöjä niin verkkosivustoa lukemalla kuin piirtämällä aineiston viestimäärien kertymistä selventäviä aikajanoja ja taulukoimalla eri alipalstojen viestimääriä (Lagus ym.

---

<sup>2</sup> <https://korp.csc.fi/>

2016: 21–32, 37).

Seuraavaksi toteutin LDA-analyysiohjelman Python-ohjelmointikielellä. Ohjelman kehitysvaiheessa käytin harjoitteluaineistoa. Tarkoituksena oli taata tutkimuksen validiteetti eli varmistaa, että toteutettu ohjelma mittaa sitä, mitä sen odotetaan mittaavan. Kun käyttää harjoitusaineistoa varsinaisen aineiston sijaan, ei tule vahvistaneeksi omaa ennakkokäsitystään siitä, mitä aineistosta ”pitäisi” löytyä. Harjoitteluaineiston käyttö tulkitsevaa analyysiohjelmalla kehitettäessä helpottaa siis tasapainottelua sensitiivisyyden ja spesifisyyden välillä. Ohjelman kehitysvaiheen lopuksi testasin sitä pienellä Suomi24-materiaalista poimitulla lemmatulla eli perusmuotoistetulla esimerkkiaineistolla<sup>3</sup>. Ohjelmoinnissa käytetty esimerkkiaineisto kerättiin Kielipankin FIN-CLARIN Korp-API:n<sup>4</sup> kautta poimimalla Suomi24-aineistosta kaikki avainsanan ”diabetes” sisältävät lauseet ja tallentamalla ne CSV-/XLS-tiedostoksi. Tämä haku tuotti tulokseksi 31 675 lausetta. Taulukossa 2 esitetään yksi näin syntyneen aineistotaulukon esimerkkirivi. LDA-ohjelma analysoi vain sarakkeen lemmas sisältämän tekstin.

Kolmannessa vaiheessa poimin FIN-CLARIN Korp-API:sta koko Terveys-palstan keskustelut. Koska ennakoin aineiston olevan suuruusluokaltaan pari miljoonaa riviä, oletin aineistonkeruun kestävän kauan. Ratkaisin ongelman käyttämällä aineiston lataukseen iOpus iMacro for Firefox -työkalua<sup>5</sup>, jota ajettiin öisin 23.6.–6.7.2016 välisenä aikana. Tämä haku tuotti 1 284 tiedostoa (13 GB dataa, eli 13 000 000 000 tavua ja arviolta 2,2 miljoonaa sivua<sup>6</sup>). Korp-API:n ominaisuuksien vuoksi kerätty aineisto oli vielä puhdistettava toistuvista riveistä eli

Taulukko 2. Esimerkki FIN-CLARIN Korp-API:n kautta saatavasta aineistosta. Tokens-sarake sisältää alkuperäisen kommenttilauseen, lemmas-sarake perusmuotoistetun lauseen.

tokens	lemmas	sentence_id	text_date from	text_time from	text_discussionarea	text_subsections	text_tid	text_cid	date	total hits	params
Mun eno kuoli diabetekseen ja toisella enolla alkaa ole sama tilanne , mun tati käyttää kans piikkei mut ei oo nyt nii pahana .	mä eno kuolla diabetes ja toinen eno alkaa ole sama tilanne , mä tati käyttää kans piikkei mut ei olla nyt nii paha .	16419	20150605	180400	Terveys	Sairaudet > Diabetes	13632757	unspecified	8.6.2016 10:41	3454	corpus=S24_001; cq=[lemma='diabetes,*' %c & _text_discussionarea="Terveys.*"]; defaultcontext=1 sentence; defaultwithin=[none]; sort=none; start=0; end=9999

<sup>3</sup> Lemmaus eli lemmatisointi tarkoittaa kielitieteessä sanan muuttamista sen perusmuotoon. Tekstissä taivutusmuodot koostuvat sanojen vartaloista ja taivutustunnuksista, esimerkiksi (perusmuoto: vartalo-taivutustunnus) vesi: vede-n, vet-tä. Tekstin perusmuotoistamisessa teksti normalisoidaan, eli kunkin käytetyn sanan taivutettu muoto (veden) palautetaan perusmuotoon (vesi) ja esimerkiksi yhdyssanoista tunnistetaan niiden osasanat. Tämä tarkoittaa sellaisen kielen automaattista analysointia jossa on runsaasti taivutusmuotoja, kuten suomen kielessä.

<sup>4</sup> <https://www.kielipankki.fi/support/korpapi>

<sup>5</sup> <http://imacros.net/overview>

<sup>6</sup> Arvion perusteena (LexisNexis 2007)

deduplikoitava. Tätä varten ohjelmoin pienen ohjelman, joka poisti toistuvat rivit ja yhdisti tulostiedostot yhdeksäksi tiedostoksi. CSV-/XLS-tiedostojen koko vaihteli lopulta välillä 32–188 MB.

Käyttämäni aineiston kokoa voi kuvata vertaamalla sitä yhteiskuntatieteellisessä tietoaarkistossa saatavilla olevaan ”Suomalaisten ruokapäiväkirjat 2013” -aineistoon, joka sisältää 486 kirjoitusta TXT- ja HTML-tiedostoina (Suomalaisen Kirjallisuuden Seura SKS ja Talonpoikaiskulttuurisäätiö 2015). Näiden kirjoitusten pituus on 1–17 A4-sivua, jolloin Aila-palvelusta ladattuna aineisto on kooltaan noin 8 MB eli 8 000 000 tavua ja arviolta 680 sivua<sup>7</sup>. Ruokapäiväkirja-aineiston koko on siis noin 0,06 % Suomi24:n Terveys-palstan aineistosta. Toisin päin ilmaistuna, Suomi24 sivuston Terveys-palstan aineisto on 1 625 kertaa suurempi kuin ruokapäiväkirja-aineisto. Tämän perusteella voi arvioida, kuinka kauan Terveys-palstan aineiston läpikäynti kestäisi perinteisellä lähiluvulla ja kuinka suuren osan luettuaan hän voisi olettaa otoksen saturoituneen laadullisesti.

## Mallin valinta

Mallin valintaan johtavan analyysin kulku oli seuraava. LDA-ohjelma sai tehtäväkseen etsiä aiheita (6, 8, 20, 30, 50 ja 150 kappaletta) deduplikoiduista lemmas-sarakkeista. Tällä tavoin iteratiivisesti hyödynnettynä odotin LDA-menetelmän löytävän sekä aiheita, jotka toistuvat aiheiden valitusta lukumäärästä riippumatta, että hierarkkisesti hienosyisemmäksi hajaantuvia keskustelunaiheita. Lopullinen 50 aiheen malli on esitetty liitteessä ja sen yhteenveto on Taulukossa 3.

Kuten edellä todettiin, aineiston LDA-aihemallinnus on tulkinnallinen siinä mielessä, että tutkijan tehtävänä on kuvata ja nimetä mallinnusohjelman löytämät aiheet. Tätä varten ohjelmoin ohjelman tulostamaan löytämiensä aiheiden avainsanalistat.

Taulukko 3. Yhteenveto 50-aiheisen mallin tuloksista

Nimetyt aiheet	Aiheiden lukumäärä
Auktoriteettineuvo	2
Fraasi	2
Ihmisen osa	1
Mielenterveys	3
Narratiivi	11
Painonhallinta	5
Päihteet	2
Sairaudet	9
Vertaisneuvo	1
(Selittämättä)	14
YHTEENSÄ	50

<sup>7</sup> Arvion perusteena (LexisNexis 2007)

Analysoitujen aiheiden nimeäminen edellyttää kuitenkin myös alkuperäistekstien lähilukemista (Grimmer ja Stewart 2013). Esimerkiksi yhteiskuntatieteilijöiden usein käyttämä työkalu MALLET<sup>8</sup> ehdottaa luettavaksi aihekohtaisesti edustavimmiksi luokittelemansa dokumentit. Näihin perehtyminen tarjoaa keinon varmistua LDA-menetelmän toimivuudesta. Vaarana on kuitenkin se, että alkuperäistekstien lähilukeminen jää tähän. Koska algoritmien kyky havaita tulkinnallisuuksia on tunnetusti rajallinen verrattuna ihmiseen (Tangherlini ja Leonard 2013), on tärkeää, että ihminen perehtyy myös niihin teksteihin, joita automatisoitu menetelmä ei välttämättä pysty luokittelemaan erityisen osuvasti. Toisaalta silloin kun tutkimuskysymys kohdistuu ihmisten vuorovaikutukseen eikä tekstidokumentteihin, ei myöskään ole syytä pysähtyä mallin suorittamaan tekstidokumenttien luokitteluun. On jatkettava matkaa kohti varsinaista tutkimuksen kohdetta (vrt. Su ym. 2017).

Nämä kaksi menetelmällistä ajatusta ohjenuorana perehdyin kunkin uuden LDA-mallinnuksen (vaihe 4) jälkeen tuotettujen avainsanalistojen perusteella niin satunnaisesti valittuihin XLS-tiedostojen teksteihin (vaihe 5) kuin alkuperäisiin keskusteluihin verkossa (vaihe 6). Aiheet Painonhallinta, Fraasi, Päihteet, Sairaudet ja Mielenterveys hajautuivat aiheiden lukumäärän kasvaessa. Hajautumisen perusteella päätin, että näistä keskustelunaiheista löytyisi erilaisia keskustelun tapoja. Toisaalta Ihmisen osa ja sekä Auktoriteetti- että Vertaisneuvo pysyivät melko selkeästi omina kokonaisuuksinaan aiheiden lukumäärästä riippumatta. Tämä viittaa siihen, että nämä aiheet määräytyvät yksittäisten voimakkaasti klusteroituvien avainsanojen kautta.

Lopullista aiheiden lukumäärää valitessani (vaihe 7) järjestin kaikkien siihenastisten mallien avainsanojen tulosteet lähiluvun perusteella käsin ryhmiksi, joille katsoin voivani nimetä otsikon. Vaikeasti nimettävien aiheiden tulosteita tarkastellessani ja lähilukiessani niiden avainsanoilla paikallistettujen tekstien XLS-tekstejä sekä verkkokeskusteluja (vaiheet 5 ja 6) löysin vielä Terveysnarratiivin, joka jäsenyi useampien eri keskustelunaiheiden yhdistelmänä.

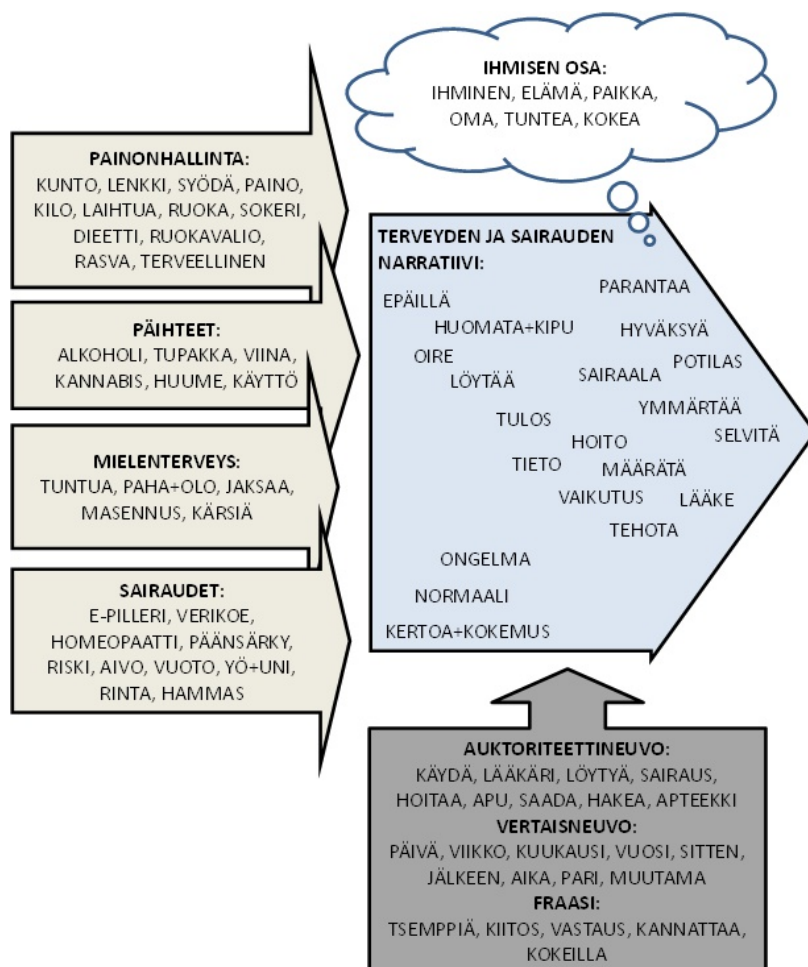
Terveysnarratiivin paikallistamisen jälkeen Taulukon 3 esittämä LDA-mallinnuksen aihekohtainen avainsanalista osoittautuu vasta analyysin välivaiheeksi, ei lopputulokseksi. Tällaiselle lähiluvun ja etäluvun lopputuloksen esittämiselle ei kuitenkaan löydy vakiintunutta muotoa. Kuvio 1 esittää yhden ehdotuksen, joka on tekstin tutkijan vapaa visualisointi. Analyysin lopputulokset käydään tämän visualisoinnin avulla yksityiskohtaisemmin läpi seuraavissa kappaleissa. Se paitsi esittelee Taulukon 3 keskustelunaiheittain valikoituja avainsanoja, myös jäsentää löydettyä narratiivia.

Aiheiden yhdistelmä sai tulkintansa analyysin palatessa vaiheesta 7 vaiheisiin 5 ja 6. Alkuperäisten verkkokeskustelujen lähiluku rohkaisi näin toteuttamaan uusia

---

<sup>8</sup> <http://mallet.cs.umass.edu/>





Kuvio 1. Analyysin tulos manuaalisesti jäsennettynä. Aihekohtaiset sanat on valittu esimerkinomaisesti.

LDA-mallinnuskierroksia (vaihe 4). Tällaisesta edestakaisesta parametrien säädöstä helposti seuraava mallin monimutkaistuminen koetaan etenkin ennustemallien tapauksessa ongelmalliseksi ylisovittamiseksi. LDA-mallinnuksenkin vaiheet voisi olettaa usein suoritettavan lineaarisesti tai yhtäaikaaisesti (esim. DiMaggio ym. 2013, liite A: 2–3, kertoo ”toteuttaneensa 8-, 9-, 10-, 11- ja 12-aiheiset mallit” ja sen jälkeen tilastollisen tarkastelun keskeisinformaatioanalyysiin). Toisaalta koneopittujen aiheiden osuvuutta koetetaan parantaa niin algoritmisesti kuin esimerkiksi tutkimusryhmässä koodaamalla validoiden (Chang ym. 2009; Grimmer ja Stewart 2013; esim. Paul ja Dredze 2014: 8). Tulkinnallisen tutkimuksen tarkoituksena ei kuitenkaan ole muodostaa ennusteita vaan ymmärtää kohteen merkityksiä ja tarkoituksia. Ilman edestakaista, lähilukuun perustuvaa tulkintaa aineiston narratiivinen näkökulma olisi jäänyt löytämättä.

## Suosituimmat keskustelunaiheet

LDA-aihemallinnusohjelma (vaihe 4) löysi koko aineistosta Taulukossa 1 esitetyistä alipalstojen aiheista suurimmat neljä: Päihteet, Mielenterveys, Sairaudet sekä Painonhallinta (ks. Kuvio 1).

LDA-menetelmällä voidaan siis löytää sosiaalisen median aineistosta suosituimmat aiheet. Kiinnostava jatkokysymys on se, millä tavoin palstalla keskustellaan näistä aiheista. Pelkkä aihekohtainen avainsanalista ei tarjoa materiaalia kysymyksen selvittämiseen. Perusmuotoisten avainsanojen listaa voi hyödyntää hakusanoina, joiden avulla palataan alkuperäisiin tekstin lauseisiin (vaihe 5). XLS-tiedostoista satunnaisesti etsimällä voi näin paikallistaa varsinaisen viestin lisäksi myös ympäröivien viestien kirjoittajien nimimerkkejä, keskustelunaiheiden otsikoita sekä muuta kirjoitusten kontekstia. Esimerkiksi kokonaisia fraaseja käyttämällä voi toteuttaa vielä tarkemman haun verkon Suomi24-palstalle ja paikallistaa vuorovaiikutuksellisia keskusteluketjuja (vaihe 6).

Alla olevassa viestissä kirjoittaja kysyy muilta kävijöiltä kokemuksia erään diabetes-lääkkeen käytöstä. Viesti on poimittu miljoonien joukosta etsimällä ensin Taulukon 2 muotoisesta, satunnaisesti valitusta XLS-tiedostosta Excelin hakutoiminnolla keskusteluja, joissa esiintyi löydetyn "Sairaudet"-aiheen avainsanoja (vaihe 5), ja sen jälkeen etsimällä keskustelufraasien avulla vastaavantyyppinen kokonainen keskustelu Suomi24.fi-verkkopalstalta (vaihe 6). Hyvää tutkimusetiikkaa noudattaen valitsin esimerkkinä käytettäväksi viestin, joka ei esimerkiksi ollut itsensä alaikäiseksi määrittäneen kirjoitus tai Mielenterveys-palstan henkilökohtaisia tuntemuksia vuodattava viesti:

"Tänään minulle ehdotettiin ensimmäistä kertaa 16 vuoden diabetes-ajan jälkeen, että siirtyisin Lantus-nimiselle insuliinille. Olen aina ennen käyttänyt erityyppisiä lyhyt- ja pitkävaikutteisia, enkä näin ollen tiedä, kuinka suhtautuisin asiaan. Lantuksen esitteessäkin luetellut edut kuulostavat lupaavilta, mutta on ymmärrettävää, että siinä ei todellisia ongelmia kerrota suoraan."

"Näin siis kysyisin teiltä jotka ovat jo jossain vaiheessa siirtyneet Lantuksen käyttäjiksi, onko siitä kokemuksia, hyviä tai huonoja? Mitä minun kannattaisi odottaa jos päätän vaihtaa, vai kannattaisiko hommaa pistää edes harkintaan? Itse tietenkin keskustelen pitkään ja hartaasti lääkäriäni kanssa, mutta haluaisin myös kuulla "normaalin elämän" tuomia kokemuksia kyseisestä mönjästä. ---"<sup>9</sup>

Vaiheiden 5 ja 6 siirtymän eli laadullisen loikan ansiosta on näin päästy käsiksi yksittäisiin viesteihin verkkokontekstissaan. Esimerkkiviestissä hoitavan lääkärin ohjeiden täydennykseksi etsitään vertaistukea ja -kokemuksia (Ahmad ym. 2006; Prior 2003). Tällä tavoin LDA-menetelmän avustamana Suomi24-aineistosta voisi kerätä viestejä ja kokonaisia viestiketjuja sisältävän aineiston esimerkiksi vertaistuen tutkimusta varten, tai koettaa tavoitella potilaskuluttajien (Autio ym. 2012) käsityksiä lääkärin 2000-luvun roolista suhteessa vertaistukeen ja verkosta saatavaan terveystietoon.

---

<sup>9</sup> Nimimerkki "TT3", 13.10.2005, palsta "Terveys > Sairaudet > Diabetes > lantukseen siirtyminen", <http://keskustelu.suomi24.fi/t/2171953/lantukseen-siirtyminen>

## **Palstarajat ylittävä keskustelunaihe: Ihmisen osa**

Ennakko-odotusten mukaisten, suosituimpien terveyskeskustelunaiheiden lisäksi LDA-menetelmä (vaihe 4) paikallisti keskustelunaiheen, jonka nimesin "Ihmisen osan" pohdinnaksi (Kuvio 1). Esimerkiksi seuraavassa viestissä tunnustetaan ihmisen rikkinäisyys ja vajavaisuus.

"Hyvin sanottu itseltäsi tuo viimeinen lause. Useimmat meistä ovat jollain tavoin rikkoontuneita tai virheellisiä ainakin, jos ei rikkoontuneita. Kaikilla meillä on kipeät kohtamme ja virheitäkin teemme."

"Mutta riittävän eheä ihminen välttää tahallisesti kiusaamista muita ihmisiä ja riittävän eheä ihminen ei värvää porukkaa tahallisesti kiusaamaan ketään yksittäistä ihmistä."<sup>10</sup>

XLS-tiedostojen lähiluvun (vaihe 5) perusteella selvisi, että tätä keskustelua käydään esimerkiksi Kuolema ja suru -palstalla, mutta myös Henkinen hyvinvointi ja mielenterveys -palstalla sekä eri sairauksien palstoilla. Tutkimalla Suomi24.fi-verkkosivustoa (vaihe 6) selvisi, että samaa aihepiiriä käsitellään Terveys-palstan lisäksi muuallakin Suomi24.fi-keskusteluissa, esimerkiksi Ajanviete-, Yhteiskunta- ja Nuoret-alipalstoilla, joita ei ollut poimittu mukaan analysoituun XLS-aineistoon.

LDA-menetelmä paikallisti näin keskustelunaiheen, jota pohditaan useam-malla kuin yhdellä Suomi24-palstalla. Tämä vahvisti ennakkoepäilyksen siitä, että sivuston aihejaotteluista huolimatta siellä keskustellaan myös keskustelualueiden rajat ylittävistä aiheista. Jos tutkija haluaa perehtyä tämän tyyppiseen tematiikkaan, ilmiö tulisikin ottaa huomioon pohdittaessa aineiston rajauksia.

## **Fraasien ympärille kiertyvät kehykset**

Kolmas LDA-menetelmän paikallistama aihetyyppi on eräänlainen tyyppilinen keskustelun tapa, jonka voidaan käsittää kiertyvän toistuvien fraasien ympärille. Vaiheiden 4, 5 ja 6 suorittamisen jälkeen nimesin nämä fraasit seuraavasti: "Auktoriteettineuvo", "Vertaisneuvo" ja "Fraasi" (ks. Kuvio 1).

"Kasvisruoka ja metformiini ovat sellainen yhdistelmä että tarvitsee joskus B12 lääkityksen avukseen. Lievemässä tilanteessa riittää re-septivapaat pillerit, vakavammassa tilanteessa 3 kk välein piikitettävä reseptilääke. Kannattaa ehdottomasti otattaa labrakoe muutaman kuukauden kuluttua kun B12 varasto alkaa hupenemaan."

---

<sup>10</sup> Nimimerkki "siksi", 25.5.2015, palsta "Terveys > Henkinen hyvinvointi ja mielen-terveys > Yleistä mielenterveydestä", <http://keskustelu.suomi24.fi/t/13609285/mika-saa-ihmisen-vainoamaan>

"Metformiinilääkkeiden ohjeissa on mainittu sivuvaikutuksena B12 vi-tamiinin alenema. Hoitavan lääkärin tulisi huomioida asia varsinkin jos tietää potilaan noudattavan kasvisruokavaliota. Normaalisti mm. liharuuat korjaavat B12 vita-miinin vähenemisen metformiininlääkityksellä."<sup>11</sup>

"Auktoriteettineuvo"-aiheeseen lukeutuvassa viestissä ohjataan kansakeskustelijaa lääkärin pakeille "otattamaan" laboratoriokoe, mutta tämän neuvon lisäksi keskustelijalle tarjotaan tietävältä ja asiantuntevalta vaikuttavia yksityiskohtaisia ohjeita.

"Kiitos tosi paljon vastauksesta! Kuulostaa erittäin tutulta! Minullakaan ei ole löydetty mitään poikkeavaa esim. verikokeista, sydänfilmistä tai keuhkokuvista. Sydämen ultraäänitutkimus ja holter vielä tekemättä."

"Minulla myös pahimmillaan tosi heikko olo, vaikka vain tiskasi. Be-tasalpaajia syön koko ajan tarvittaessa, nyt noin kerran päivässä. Yöt ovat pahimpia, kun sydän ei aina rauhoitu päivän rasituksesta vaan pumpppaa voimakkaasti, joskus myös nopeasti. Onneksi betat tehoavat."

"Sairastelu on nyt kestänyt n. 6 viikkoa enkä näillä näkymin ole vähään aikaan pääsemässä liikkumaan. Minkälaisia oireita sinulla on nyt? Tykyttelyä, uupumista vai mitä?"<sup>12</sup>

Toisaalta keskustelupalstalla jaetaan kuluttajien lääkäriin ohjaamisen lisäksi terveysneuvoja tasa-arvoisemmalta pohjalta, kuten yllä "Vertaisneuvo"-aiheeseen lukeutuvassa esimerkissä. Siinä kirjoittaja tarjoaa ja etsii vertaistukea kuvailemalla omia sydänoireitaan, joihin ei ole vielä saatu lopullista diagnoosia mutta joihin lääkitys on jo silti auttanut. Viesti sisältää myös "Fraasi"-aiheelle tyypilliset kiitokset kansakeskustelijoille.

## **Terveyden ja sairauden narratiivi**

Neljäs löydös oli Terveyden ja sairauden narratiivi (ks. Kuvio 1), joka löytyi analyysin viimeisessä, aiheiden lukumäärän iterointivaiheessa (vaihe 7). Ohjelmiston mallintaessa yhä suurempaa aiheiden lukumäärää aiheiden avainsanat ja niille luokemalla hahmotetut kuvaukset alkoivat hajaantua siinä määrin, että päätin jäsentää aiheita toisin kuin vain staattisina sisältökategorioina. Edellä kuvattu avainsanatuloisteiden järjestäminen käsin yhdistettynä XLS-aineistoon (vaihe 5) ja alkuperäisten verkkokeskustelujen lähilukemiseen (vaihe 6) tuottikin hahmotelman

---

<sup>11</sup> Nimimerkki "labrat välillä", 21.12.2012, palsta "Terveys > Sairaudet > Diabetes", <http://keskustelu.suomi24.fi/t/11133702/kasvisruoka->

<sup>12</sup> Nimimerkki "Apua kaipaava", 1.4.2008, palsta "Terveys > Sairaudet > Sisätaudit > Sydän- ja verisuonitaudit", <http://keskustelu.suomi24.fi/t/5546632/lieva-sydänlihastulehdus>

yleisestä narratiivista, jolla sairauksia jäsennetään kirjoittajien keskuudessa.

Aiemmassa tutkimuksessa narratiivit on tunnistettu osaksi kulttuurista terveyttä ja yksilöiden selviytymistä (Charon 2001; Kalitzkus ja Matthiessen 2009; Karjalainen ja Kotkavirta 2001; Leimumäki 2012; Puuronen 2012). Narratiivisuus tarkoittaa tällöin potilaiden omaa sairautta koskettavia yksilöllisiä kertomuksia, kokemusten vaihtamista toisten kanssa. Kokemusten jakamiseen rohkaisee keskustelijoiden tapa pyytää toisiltaan mielipiteitä ja kokemuksia, kuten alla oleva esimerkki osoittaa:

"Pari kuukautta ollut vaivana sydämen välilyönnit, eli tunne siitä miten sydän jättää lyönnin välistä. Vaiva on ollut minulla olemassa niin kauan kuin muistan, (ikäni 27v ja mies) mutta ennen se on ollut satunnaista kun nyt taas välejä saattaa tulla vaikka minuutin välein. Syke on muutenkin hieman heitellyt ja ollut tavallista hermostuneempi, kuten olokin."

"Lääkäri on kuunnellut pariin otteeseen sydäntä, on otettu verikokeita ja EKG, mutta mitään ihmeellistä ei ole havaittu. Syyksi lääkäri epäili stressiä, joka on hyvin mahdollista, ja lääkkeeksi rentoutumista sekä tarvittaessa Propral 10mg."

"Ymmärtääkseni kyseistä lääkettä otetaan esim. esiintymispelkoon, rauhoittamaan sykettä."

"Hieman jännittää ottaa pilleri, jos se sykkeeseen vaikuttaa. Pysähtyy sydän vielä lopullisesti? Mitä mieltä olette kuinka Propral sopii minun kaltaiselleni välilyöntitapaukselle?"<sup>13</sup>

Suomi24-kirjoitusten narratiiveissa ilmenee koko sairastumisen kaari epäi-lystä pelkojen kautta diagnoosiin ja mahdolliseen parantumiseen tai selviytymiseen. Narratiivisuus on keskeinen osa terveyttä ja yksilöiden selviytymistä, eikä narratiivin löytyminen sinänsä ole yllättävää. Suomi24-keskustelut ja niihin osallistuminen auttavat myös osaltaan kirjoittajia ja lukijoita heidän sopeutumisprosessissaan. Verkkokeskustelun automatisoidun analyysin yhdistäminen laadulliseen loikkaan tarjoaa kuitenkin keinon tarkastella narratiiveja sisältävää yhteisön vuorovaikutusta uudella tavalla. Yksilöllistenkin tarinoiden rakentumisella on lainalaisuuksia, jotka ovat osin riippumattomia yksilön todellisuudesta (Karjalainen ja Kotkavirta 2001). Kun jaetut yksilöllisten persoonien henkilökohtaiset tarinat muotoutuvat kirjoittajien ja lukijoiden yhteistyönä yhteiseksi ja julkiseksi merkityksiä sisältäväksi kertomukseksi, ne muodostavat sosiokulttuurisia terveyden ja sairauden metanarratiiveja (Kalitzkus ja Matthiessen 2009).

---

<sup>13</sup> Nimimerkki "Minä vaan", 25.11.2008, palsta "Terveys > lääkkeet", <http://keskustelu.suomi24.fi/t/6525014/propral-ja-valilyonnit>

Esimerkiksi erään keskusteluketjun aloittaja epäilee äitinsä saamaa muistisairausdiagnoosia ja kyselee, voisivatko tutkimustulokset kenties olla väärä. Aloituskeskusteluun vastaavassa viestissä alla ilmenee potilaan ja hoitohenkilökunnan välisen vuorovaikutuksen merkitys (Kalitzkus ja Matthiessen 2009). Vastaaja kertoo, kuinka heidän tapauksessaan hoitaja on tehnyt liian ”oma-aloitteisen päätelmän” ja siten ”hankaloittanut asioita”, mutta ”nuoren” lääkärin kyky ”myöntää” rajallinen tietomääränsä on ”onneksi” edistänyt asiaa. Kirjoittaja muodostaa muistisairaiden omaisille metanarratiivista tarinaa, jossa oikean diagnoosin löytymisessä sinnikkyys kyllä palkitaan, mutta myös ”silkkaa tuuria” vaaditaan.

"Älkää ihmeessä jättäkö asiaa siihen jos terveyskeskuksen muistitutkimuksissa ei mitään selvinnyt ja te omaiset kuitenkin tiedätte että jotain on pielessä. Meillä jo muistitestin tehnyt hoitaja ilmoitti oma-aloitteisesti potilaalle ettei dementiaa ole ja hankaloitti asioita huomattavasti. Onneksi lääkäri oli nuori ja myönsi ettei tunne eri dementiaa riittävästi ja lähetti jatkotutkimuksiin. Päänkuvausta, Cerad testia, neurologia, psykologia ja geriatria - monta viisasta päätä lopulta löytyi ja syy kaikkiin muutoksiin frontaalidementia. Vaikka tieto lisää tuskaa sanotaan antoi se meille kuitenkin viitteet miten toimia jatkossa - silloin emme arvanneet että jatkossa tauti eteni isoin harppauksin ja oli onni että olimme ajoissa liikenteessä. No asiat nyt hyvin mutta omaisilta vaatii aika sinnikkyyttä ja välillä silkkaa tuuria - niinhän ei saisi olla! Lääkkeiden anto on kyllä niin ammattijuttu että kannattaisi miettiä sitä ihan potilasturvan kannalta - tai mitä sitten tarkoittikin omasta päätöksestä en tiedä... Neuvoja varmaan satelee mutta kehoittaisin vaatimaan läheteitä eteenpäin."<sup>14</sup>

Kun vakavan sairauden vuoksi esimerkiksi oma aikaisempi elämäntapa täytyy muuttaa, sairauden kanssa ”sinuiksi pääseminen” vie jopa vuosia. Vaikka kaikista sairauksista, kuten kroonisesta reumasta, ei voikaan parantua, tarinoista ilmenee onnistuneita oman kokemuksellisen identiteetin jatkuvuuden kokemuksia (Karjalainen ja Kotkavirta 2001). Toisia kirjoittajia diagnoosin löytäminen ja asioiden selkeytyminen helpottaa, jolloin identiteetin löytyminen ilmenee narratiivissa. Kaikkien narratiivien juonissa ei kuitenkaan sopeuduta (Leimumäki 2012). Kuten aiemmissakin keskusteluesimerkeissä ilmeni, lääkärin määräämien lääkkeiden sivuvaikutukset mietityttävät keskustelijoita. Sivuvaikutusten vuoksi lääkkeistä halutaan toisinaan luopua kokonaan, kuten alla olevassa viestissä. Ratkaisua pohtiessaan kirjoittaja on ensin pyytänyt kokemuksia muilta luopujilta ja jakaa sitten kuvauksen lopettamisen seurauksista.

"Olen viestiketjun aloittaja ja haluan kertoa mitä todella tapahtui, kun lopetin lääkkeet. Aikomukseni oli lopettaa lääkkeet vähitellen, mutta diagnoosiini ja luonteeseni sopien, lopetin ne kerralla. Ensimmäiset viikot huimasi ja koko ajan

---

<sup>14</sup> Nimimerkki ”Seela”, 31.8.2006, palsta ”Terveys > Sairaudet > Neurologiset sairaudet > Dementia”, <http://keskustelu.suomi24.fi/t/3340554/ei-diagnoosia>

oli kamala vesijano.Noin kolmen viikon päästä lopetuksesta ajatukset alkoivat kirkastua, jaksoin lähteä ulos, kykenin tekemään päätöksiä ym. ym. jaksoin aloittaa liikkumisen ja sen myötä vähensin syömistä.Painoni on pudonnut lähes 20kg.Olen voinut palata työhön, jota en koskaan uskonut kykeneväni enää tekemään.Koin uuden lääkkeettömän elämäni alkuaikoina, että teen jotain väärää, kun en syö minulle määrättyjä lääkkeitä.Lopulta ajatukset muuttuivat ja päätin toimia niin kuin parhaaksi näin.En yllytä ketään tekemään samoin kuin minä.Jollekin lääkkeiden yhtäkkinen lo-pettaminen voi olla huono ratkaisu ja varmaan aina olisi järkevää keskustella asiasta lääkärin kanssa.Ihanaa kevättä teille kaikille lääkkeillä tai ilman!"<sup>15</sup>

Lääkkeiden käytön lopetuksen tyytyväinen kuvaus muodostaa julkisesti sosiokulttuurista narratiivia siitä, että lopettaminen voi toimia yksilöllisesti varsin hyvin. Lopettajalla saattaakin olla kirjoitushetkellä henkilökohtainen tuntemus onnistuneesta päätöksestä, mutta Suomi24-palstan avoimuus antaa muille kirjoittajille tilaa tarjota sosiokulttuuriseen narratiiviin täydentävä kertomus. Omatoiminen lääkkeiden käytön lopetus saattaa aiheuttaa vaikeuksia, joista sekä potilas itse että tämän lähipiiri kärsivät, kuten esimerkkiviestissä alla:

"Vaikeista maanisista kausista kärsivän ihmisen lähiomaisena suorastaan kyr\*ii lukea tätä ketjua. Toimitin juuri hoitoon oman äitini, joka omatoimisesti lopetti Zyprexan käytön ja nousi taas maaniseen vaiheeseen, jona aikana uhkaili mm. sukulaisia hengenlähdöllä ja sai itsensä tolkuttomaan taloudelliseen kuseen. Itse suhtaudun asiaan vast' edes niin, että kun ja jos äitini seuraavan kerran ryhtyy itsekseen rukkaamaan lääkityksiään, niin minun puolestani hän voi tappaa itsensä vaikka samantien. Niin paljon hän on minunkin perheeni elämää rasittanut, siis omien lastensa ja lastenlastensa elämää. Lääkityksen omatoiminen rassaaminen on merkki silkasta välinpitämättömyydestä muita kohtaan. Sitä en ihmettele yhtään, että elämä tuntuu tasapaksulta lääkityksessä: sitähan se normaali elämä ihmisellä on. Tämä asia ei vain bipoille valkene, sillä ovat niin mieltyneet niihin epärealistisiin manioihinsa ja toisten ihmisten kiusaamiseen."<sup>16</sup>

Yhteenvetona, yhdistämällä LDA-koneoppimismenetelmä ja laadullinen loikka Terveys-palstan aineistosta löytyi keskustelunaiheiden lisäksi yksilöllisiä terveyden ja sairauden narratiiveja. Keskustelupalstan avoimuuden vuoksi Suomi24-keskusteluihin kietoutuu paitsi potilaiden, myös lähipiirin ja omaisten täydentäviä narratiiveja. Näiden tutkiminen paljastaa merkkejä kansalaiskeskusteluissa ilmenevästä sairauden sosiokulttuurisesta narratiivista. Tämän tyyppisille

---

<sup>15</sup> Nimimerkki "alkuperä", 8.5.2015, palsta "Terveys > Henkinen hyvinvointi ja mielen-terveys > Kaksisuuntainen mielialahäiriö", <http://keskustelu.suomi24.fi/t/11333992/laakkeiden-lopettaminen!>

<sup>16</sup> Nimimerkki "Vtunväsynyt", 23.10.2015, palsta "Terveys > Henkinen hyvinvointi ja mielen-terveys > Kaksisuuntainen mielialahäiriö", <http://keskustelu.suomi24.fi/t/11333992/laakkeiden-lopettaminen!>



tutkimuskysymyksille Suomi24-aineisto on epäilemättä ainutlaatuisen rikas aineisto, jonka tarkastelussa automatisoitu aineiston tiivistäminen ja laadullinen loikka voivat olla suureksi hyödyksi.

## Johtopäätökset

LDA-menetelmä löytää odotetusti Terveys-palstan suurimmat keskustelunaiheet, mutta myös useammilla alipalstoilla ilmeneviä aiheita ja fraaseja. Suomi24:n tapaista laajaa vuorovaikutuksellista aineistoa voi tarkastella vahvemmin tulkitsevilla otteella edellä kuvattuun vaiheittaiseen tapaan. Hyödyntämällä automatisoidun analyysin avainsanastoja hakusanoina pidempien tekstien löytämiseksi laadullinen loikka tuo verkon alkuperäistekstit tutkijan ulottuville.

Analysoimalla Terveys-palstaa sekä automatisoidusti että lähilukemalla aineistosta sai kattavan otteen. Olisi toki mahdollista poimia tekstien joukosta esimerkkiaineisto verkkokeskustelupalstaa läpikäymällä. On kuitenkin tärkeää hahmottaa, kuinka paljon massiivisempi verkkoaineiston sivumäärä onkaan verrattuna tyypilliseen laadullisen aineiston kokoon. Pelkästään lähiluvun keinoilla tarkastellut satunnaiset tekstit Suomi24:n verkkosivulta varmaankin tuottaisivat jonkin kiinnostavan näkökulman. Terveys-palstan laajuudesta johtuen kokonaiskuvan hahmottaminen olisi kuitenkin vaikeaa. Jokin olennainen näkökulman täydennys joltain tietyltä ajanjaksolta saattaisi jäädä huomiotta.

Toinen esimerkki laadullisen loikan hyödyllisyydestä on terveyden narratiivin paikallistaminen. Alkuperäisen verkkokeskustelun tulkitseminen LDA-ohjelman avainsanoja hyödyntäen paljasti niin yksilöllisiä kuin sosiokulttuurisia narratiiveja. Kirjoitushetkellä narratiivien automatisoitu tutkimus ei näytä olevan vielä vakiintunutta, ja sitä puhtaasti automatisoidusti kartoittavien menetelmien kehittyminen on satunnaista ja yksittäisten tutkijoiden varassa (Hossain ym. 2012; Kumar ym. 2015; myös Mohr ja Bogdanov 2013: 559). Tulevaisuudessa menetelmät voivat kehittyä, mutta tulkintojen muodostamisessa tarvittaneen jatkossakin lähiluvun ja etäluvun yhteispeliä.

Selkein analyysin puute ilmenee aineiston tekstidokumenttiyksikön valinnassa. Tutkimusyksiköksi valikoitui sekä alkuperäisaineiston että FIN-CLARIN:in tietokannan rakenteesta johtuen lause, joka voi olla varsin lyhytkin tutkimuskohde. Kielipankki ja Suomen Akatemian Citizen Mindscapes projekti kehittävätkin parhaillaan Suomi24-aineiston tietokantoja ja hyödyntämistyökaluja. Pyrkimyksenä on, että aineistoa on tulevaisuudessa mahdollista analysoida etälukemalla myös viestien ja keskustelujen tasolla. Tämän voi olettaa tuottavan aineistoon vielä rikkaamman näkymän.

## Aihemallinnuksen mahdollisuudet kulutustutkimuksessa

Monipuolisia aineistoja ja menetelmiä hyödyntävien kulutustutkijoiden on tärkeää



perehtyä syvällisesti myös verkkoaineistojen käsittelyyn. Tämä artikkeli kuvaa esimerkkien avulla, millä tavoin LDA-aihemallinnusta voidaan hyödyntää kaivattuna aineiston tiivistämisen menetelmänä tekstien tulkitsevan tutkimuksen tarpeisiin (Tangherlini ja Leonard 2013). Koneoppimismalleista ja XLS-aineistosta alkuperäisteksteihin palaava laadullinen loikka tuo jäntevyyttä tuhansia kertoja perinteisiä aineistoja suurempien tekstiaineistojen tutkimukseen. Etä- ja lähilukua yhdistävä tutkimusmenetelmä tarjoaa tilaa paitsi täydentävälle tutkimukselle, myös kokonaan uusille lähestymistavoille. Tutkijan arjessa tämä voi tarkoittaa niin uudentyyppisiä tutkimuskysymyksiä kuin uudentyyppisten tutkimusryhmien muotoutumista, joissa ohjelmoinnin osaaaja työskentelee rinta rinnan tulkitsevan tutkijan kanssa. Useat artikkelissa käsitellyistä tutkimuksista ovatkin tällaisen yhteistyön tulosta, mm. DiMaggio ym. (2013); Gonzalez-Bailon ja Paltoglou (2015); Hannak ym. (2012); Schwartz ym. (2013).

Niin liike-elämän toimijoiden kuin tutkijoidenkin keskuudessa on jo huomattu, ettei suurten aineistojen analyysi olekaan niin helppoa tai tarkkaa kuin ensi-innostuksen keskellä oletettiin (Lazer ym. 2014). Kriitikon ohjelmistojen hyödyntäminen saattaa osoittautua ongelmalliseksi kulutustutkimuksessa ja muissa yhteiskuntatieteissä, kuten muissakin niitä käyttävissä tieteissä (Kallio 2016). On myös selvää, että muotoilemattoman tekstin lähiluku vailla ymmärrystä tekstin alkuperäisestä esityskontekstista kadottaa jotain olennaista tekstin tuottajien ja vuorovaikutuksen tulkinnoista. Suomi24-aineisto on tässä mielessä poikkeus, koska alkuperäiset keskustelut ja niiden verkkokontekstin voi löytää Google-haulla.

Suurten aineistojen analyysimenetelmien kehitys on kiivaassa vauhdissa, vaikka analytiikkainnostuksen huippu lienee jo ohitettu. Verkkovuorovaikutuksen analysointi tarjoaa edelleen paljon mahdollisuuksia tieteenaloille, joissa vuorovaikutus on tärkeä tutkimuskohde. Esimerkiksi narratiivisessa lääketieteessä hyödynnetään jo tarinoiden merkityksiä paitsi yksilön parantumisen apuna, myös lääkärin ja potilaan välisessä viestinnässä (Kalitzkus ja Matthiessen 2009; Karjalainen ja Kotkavirta 2001). Suomi24-aineiston narratiivien aihehallinnusavusteinen lähiluku voisikin tarjota tietoa sosiokulttuurisista narratiiveista ja siten tukea potilaskuluttajakeskisemmän lääketieteen edistämistä Suomessa.

Etäluvun avulla hankituista uusista tulkinnoista voi ajan myötä aueta uusia teoreettisia maisemia. Vaikka menetelmien opettelu ja niiden menestyksekkäs hyödyntäminen vaativatkin aikaa ja vaivaa, menetelmät tuottavat uusia mahdollisuuksia tekstejä tulkitsevalle kuluttajatutkimukselle. Verkkoaineistojen jatkuva kertyminen mahdollistaa kuluttajien arkipäivän dynamiikan ja markkinoiden toiminnan ymmärtämisen entistä tarkemmin ja syvällisemmin.

## Kiitokset

Kiitän kaikkia artikkeliani kommentoineita anonyymeja vertaisarvioitsijoita, Kuluttajatutkimuskeskuksen tutkijoita, etenkin Mika Pantzaria, Minna Ruckensteinia

ja Mari Nivaa, sekä muita eri vaiheissa palautetta antaneita, erityisesti Mikko Mattilaa, Mikko Salasuota sekä Helsingin yliopiston talous- ja sosiaalihistorian tohtoriseminaarin osallistujia.

## Lähteet

- Ahmad, Farah, Hudak, Pamela L., Bercovitz, Kim, Hollenberg, Elisa & Levinson, Wendy (2006) 'Are physicians ready for patients with Internet-based health information?', *Journal of Medical Internet Research* 8(3): 1–11.
- Ahonen, Pertti (2015) 'Institutionalizing Big Data methods in social and political research', *Big Data & Society*, 2(2): 16 s.
- Autio, Minna, Helovuori, Susanna & Autio, Jaakko (2012) 'Potilaskuluttajan ja lääkärin muuttuvat roolit sähköistyvillä terveystietomarkkinoilla', *Kulutustutkimus. Nyt!*, 6(2): 40–57.
- Berry, David M. (2011) 'The computational turn: Thinking about the digital humanities', *Culture Machine*, 12: 1–22.
- Blei, David (2012) 'Probabilistic topic models', *Communications of the ACM*, 54(4): 77–84.
- Buurma, Rachel Sagner (2015) 'The fictionality of topic modeling: Machine reading Anthony Trollope's Barsetshire series', *Big Data & Society*, 2(2): 6 s.
- Centola, Damon (2013) 'Social media and the science of health behavior', *Circulation*, 127(21): 2135–44.
- Chang, Jonathan, Boyd-Graber, Jordan L., Gerrish, Sean, Wang, Chong & Blei, David M. (2009) 'Reading Tea Leaves: How Humans Interpret Topic Models', teoksessa *Advances in Neural Information Processing Systems 22 (NIPS 2009)*, Vancouver.
- Charon, Rita (2001) 'Narrative Medicine: A model for empathy, reflection, profession, and trust', *Journal of American Medical Association*, 286(15): 1897–1902.
- DiMaggio, Paul, Nag, Manish & Blei, David (2013) 'Exploiting affinities between topic modeling and the sociological perspective on culture: Application to newspaper coverage of U.S. government arts funding', *Poetics*, 41(6): 570–606.
- Dolbec, Pierre-Yann & Fischer, Eileen (2015) 'Refashioning a Field? Connected Consumers and Institutional Dynamics in Markets', *Journal of Consumer Research*, 41(6): 1447–68.
- Gil de Zúñiga, Homero & Diehl, Trevor (2017) 'Citizenship, Social Media, and Big Data: Current and Future Research in the Social Sciences', *Social Science Computer Review*, 35(1): 3–9.
- Gonzalez-Bailon, Sandra & Paltoglou, Georgios (2015) 'Signals of Public Opinion in Online Communication: A Comparison of Methods and Data Sources', *ANNALS of the American Academy of Political and Social Science*, 659(1): 95–107.
- Grimmer, Justin & Stewart, Brandon M. (2013) 'Text as data: The promise and pitfalls of automatic content analysis methods for political texts', *Political Analysis*, 21(3): 267–97.
- Guggenheim, Lauren, Jang, S. Mo, Bae, Soo Young & Neuman, W. Russell (2015) 'The Dynamics of Issue Frame Competition in Traditional and Social Media', *ANNALS of the American Academy of Political and Social Science*, 659(1): 207–24.
- Hannak, Aniko, Anderson, Eric, Feldman Barrett, Lisa, Lehmann, Sune, Mislove, Alan & Riedewald, Mirek (2010) 'Tweetin' in the Rain: Exploring Societal-

- scale Effects of Weather on Mood', teoksessa Sixth International AAAI Conference on Web and Social Media (ICWSM-12), Dublin.
- Hardey, Michael (2001) 'E-health': The Internet and the transformation of patients into consumers and producers of health knowledge', *Information, Communication & Society*, 4(3): 388–405.
- Haverinen, Anna & Suominen, Jaakko (2015) 'Koodaamisen ja kirjoittamisen vuoropuhelu? – Mitä on digitaalinen humanistinen tutkimus', *En-nen ja nyt – historian tietosanomat*.
- Heinonen, Visa, Mykkänen, Juri, Pantzar, Mika & Roponen, Seppo (1996) 'Suomalaisen talouspolitiikan ajattelumallit valtiovarainministerien budjettiesitelmissä 1974–1994'. Helsinki: Kuluttajatutkimuskeskus.
- Hossain, M. Shahriar, Gresock, Joseph, Edmonds, Yvette, Helm, Richard, Potts, Malcolm & Ramakrishnan, Naren (2012) 'Connecting the dots between PubMed abstracts', *PLoS ONE*, 7(1): e29509.
- Huovila, Janne & Saikkonen, Sampsa (2015) 'Establishing credibility, constructing understanding: The epistemic struggle over healthy eating in the Finnish dietetic blogosphere', *Health*, 20(4): 383–400.
- Jauho, Mikko (2016) 'The social construction of competence: Conceptions of science and expertise among proponents of the low-carbohydrate high-fat diet in Finland', *Public Understanding of Science*, 25(3): 332–45.
- Kalitzkus, Vera & Matthiessen, Peter F. (2009) 'Narrative-based medicine: Potential, pitfalls, and practice.' *The Permanente journal*, 13(1): 80–86.
- Kallio, Aleks (2016) 'Properties of Fixed-Fixed Models and Alternatives in Presence-Absence Data Analysis', *PLoS ONE*, 11(11): e0165456.
- Karjalainen, Kerkko & Kotkavirta, Jussi (2001) 'Narratiivinen lääketiede Mitä se on ja miksi se on tärkeää?', *Suomen lääkärilehti*, 5138–41.
- Kumar, Nilay, Khunger, Monica, Gupta, Arjun & Garg, Neetika (2015) 'A content analysis of smartphone-based applications for hypertension management', *Journal of the American Society of Hypertension*, 9(2): 130–36.
- Lagus, Krista, Pantzar, Mika, Ruckenstein, Minna & Ylisiurua, Marjoriikka (2016) 'Suomi24 – muodonantaa aineistolle'. Helsinki: Helsingin yliopiston valtiotieteellisen tiedekunnan julkaisuja 10.
- Lazer, David, Kennedy, Ryan, King, Gary & Vespignani, Alessandro (2014) 'The Parable of Google Flu: Traps in Big Data Analysis', *Science*, 343(14 March): 1203–5.
- Leimumäki, Anna (2012) 'Kuinka kokemuksesta voi kertoa? Sepelvaltimo-tautiin sairastuminen omaelämäkerrallisena kertomuksena', teoksessa Honkasalo, M.-L. & Salmi, H. Terveyttä kulttuurin ehdoilla Näkökulmia kulttuuriseen terveystutkimukseen. Turku: k&h, 358–96.
- LexisNexis (2007) 'How Many Pages in a Gigabyte? Discovery Services – Fact Sheet 2'. Noudettu 20.12.2016. ([https://www.lexisnexis.com/applieddiscovery/lawlibrary/whitePapers/ADI\\_FS\\_PagesInAGigabyte.pdf](https://www.lexisnexis.com/applieddiscovery/lawlibrary/whitePapers/ADI_FS_PagesInAGigabyte.pdf))
- Mohr, John W. & Bogdanov, Petko (2013) 'Editorial introduction – Topic models: What they are and why they matter', *Poetics*, 41(6): 545–569.
- Parmentier, Marie-Agnès & Fischer, Eileen (2014) 'Things fall apart: The dynamics of brand audience dissipation', *Journal of Consumer Research*, 41(5): 1228–51.
- Parviainen, Olli, Poutanen, Petro, Laaksonen, Salla-Maaria & Rekola, Mikael (2012) 'Measuring the effect of social connections on political activity on Facebook', teoksessa Internet, Politics, Policy: Big Data, Big Challenges (IPP 2012), Oxford.

- Paul, Michael J. & Dredze, Mark (2014) 'Discovering Health Topics in Social Media Using Topic Models', *PLoS ONE*, 9(8):e103408.
- Prior, Lindsay (2003) 'Belief, knowledge and expertise: The emergence of the lay expert in medical sociology', *Sociology of Health & Illness*, 25(3): 41–57.
- Purhonen, Semi & Toikka, Arho (2016) "Big datan' haaste ja uudet laskennalliset tekstiaineistojen analyysimenetelmät. Esimerkkitaapauksena aihehalliianalyysi tasavallan presidenttien uudenvuodenpuheista 1935–2015', *Sosiologia*, 53(1): 6–27.
- Puuronen, Anne (2012) 'Kerronnallinen tutkimusote anoreksiaan', teoksessa Honkasalo, M.-L. & Salmi, H. *Terveyttä kulttuurin ehdoilla Näkökulmia kulttuuriseen terveystutkimukseen*. Turku: k&h, 397–438.
- Savage, Mike (2013) 'The 'Social Life of Methods': A Critical Introduction', *Theory, Culture & Society*, 30(4): 3–21.
- Schwartz, Hansen Andrew, Eichstaedt, Johannes C., Kern, Margaret L., Dziurzynski, Lukasz, Lucas, Richard E., Agrawal, Megha, Park, Gregory J., Lakshmikanth, Shrinidhi K., Jha, Sneha, Seligman, Martin E. P. & Ungar, Lyle (2013) 'Characterizing Geographic Variation in Well-Being Using Tweets', teoksessa *Seventh AAAI Conference on Weblogs and Social Media (ICWSM 2013)*, Boston.
- Su, Leona Yi-fan, Cacciatore, Michael A., Liang, Xuan, Brossard, Dominique, Scheufele, Dietram A. & Xenos, Michael A. (2017) 'Analyzing public sentiments online: Combining human- and computer-based content analysis', *Information, Communication & Society*, 20(3): 406–27.
- Suomalaisen Kirjallisuuden Seura SKS & Talonpoikauskulttuurisäätiö (2015) 'Suomalaisten ruokapäiväkirjat 2013 [sähköinen tietoaieisto]. Versio 1.0 (2015-08-25)', Noudettu 28.11.2016. (<http://urn.fi/urn:nbn:fi:fsd:T-FSD3044>)
- Tangherlini, Timothy R. & Leonard, Peter (2013) 'Trawling in the Sea of the Great Unread: Sub-corpus topic modeling and humanities research', *Poetics*, 41(6): 725–49.
- Tolonen, Mikko & Lahti, Leo (2015) 'Aatehistoria ja digitaalisten aineistojen mahdollisuudet', *Ennen ja nyt – historian tietosanommat*.
- Veer, Ekant, Ozanne, Lucie K. & Hall, C. Michael (2016) 'Sharing cathartic stories online: The internet as a means of expression following a crisis event', *Journal of Consumer Behavior*, 15(4): 314–24.
- Weng, Lilian, Menczer, Filippo & Ahn, Yong-Yeol (2013) 'Virality Prediction and Community Structure in Social Networks', *Scientific Reports*, 3: 2522.
- Wilkerson, John, Smith, David & Stramp, Nicholas (2015) 'Tracing the Flow of Policy Ideas in Legislatures: A Text Reuse Approach', *American Journal of Political Science*, 59(4): 943–56.
- Winter, Lasse & Wiberg, Matti (2016) 'Presidentin uudenvuodenpuheet: Kvantitatiivisen tekstianalyysin mahdollisuuksia', *Politiikka*, 58(1): 80–88.
- Wright, Kevin & Bell, Sally (2003) 'Health-related Support Groups on the Internet: Linking Empirical Findings to Social Support and Computer-mediated Communication Theory', *Journal of Health Psychology*, 8(1): 39–54.

## Kirjoittajatiedot

DI, VTK Marjoriikka Ylisiurua työskentelee Kuluttajatutkimuskeskuksessa tohtorikoulutettavana ja suorittaa kuluttajaekonomian jatko-opintoja Helsingin yliopiston taloustieteen laitoksella. [marjoriikka.ylisiurua\(at\)helsinki.fi](mailto:marjoriikka.ylisiurua(at)helsinki.fi).

## LIITE: LDA-malli (50 aihetta) nimettynä ja esitettynä avainsanoittain

PAINONHALLINTA #5	PAINONHALLINTA #17	PAINONHALLINTA #26	PAINONHALLINTA #40
0.082*"näin"	<b>0.110*"syödä"</b>	<b>0.047*"ruoka"</b>	0.062*"suomi"
<b>0.052*"suositella"</b>	<b>0.042*"paino"</b>	0.042*"iho"	0.055*"määrä"
<b>0.052*"helppo"</b>	<b>0.029*"kilo"</b>	0.042*"katsoa"	0.048*"tapaus"
<b>0.044*"kunto"</b>	0.027*"2"	0.035*"vaiva"	0.036*"ihme"
0.020*"ko"	0.023*"silti"	<b>0.027*"syödä"</b>	0.027*"maa"
0.019*"valita"	0.023*"päivä"	0.022*"peljätä"	0.020*"onni"
<b>0.019*"aluksi"</b>	<b>0.022*"laihtua"</b>	0.018*"netti"	0.019*"verrata"
<b>0.017*"lenkki"</b>	0.020*"5"	0.016*"parantua"	<b>0.019*"dieetti"</b>
0.016*"käydä"	0.019*"3"	<b>0.015*"sokeri"</b>	<b>0.018*"hedelmä"</b>
0.015*"tehokas"	0.018*"n."	0.014*"saada"	0.016*"myydä"
PAINONHALLINTA #46	PÄIHTEET #9	PÄIHTEET #32	IHMISEN OSA #44
<b>0.039*"rasva"</b>	0.102*"miten"	0.227*"_"	<b>0.085*"ihminen"</b>
0.036*"ostaa"	0.048*"vaikuttaa"	<b>0.068*"käyttö"</b>	<b>0.085*"elämä"</b>
<b>0.033*"ruokavalio"</b>	0.042*"miettiä"	<b>0.064*"alkoholi"</b>	0.034*"koskaan"
<b>0.028*"syödä"</b>	0.040*"voida"	<b>0.054*"kannabis"</b>	<b>0.033*"tuntea"</b>
<b>0.027*"tuote"</b>	0.027*"liikkua"	<b>0.041*"huume"</b>	<b>0.032*"paikka"</b>
<b>0.019*"terveellinen"</b>	<b>0.022*"polittaa"</b>	<b>0.027*"käyttää"</b>	<b>0.025*"tunne"</b>
0.019*"hintaa"	<b>0.017*"tupakka"</b>	<b>0.024*"viina"</b>	<b>0.023*"kokea"</b>
0.018*"hankkia"	0.016*"asia"	0.023*"kuu"	0.021*"loppu"
0.017*"suomalainen"	0.014*"saada"	0.022*"ohje"	0.017*"voida"
<b>0.016*"sisältää"</b>	0.014*"tahansa"	<b>0.015*"käyttäjä"</b>	<b>0.017*"oma"</b>
MIELENTERVEYS #0	MIELENTERVEYS #3	MIELENTERVEYS #16	VERTAISNEUVO #13
<b>0.092*"tuntua"</b>	0.127*"jo"	0.079*"jäädä"	0.103*"mennä"
<b>0.060*"paha"</b>	0.116*"vuosi"	0.036*"tärkeä"	<b>0.065*"päivä"</b>
<b>0.055*"olo"</b>	0.034*"välillä"	0.032*"raha"	<b>0.054*"jälkeen"</b>
<b>0.044*"jaksaa"</b>	0.031*"kestää"	0.026*"kautta"	<b>0.051*"viikko"</b>
0.035*"usein"	0.029*"aika"	0.022*"yleinen"	<b>0.031*"pari"</b>
0.023*"parempi"	<b>0.027*"masennus"</b>	0.021*"mahdollisuus"	<b>0.030*"sitten"</b>
0.021*"jotenkin"	<b>0.026*"kärsiä"</b>	0.020*"d"	<b>0.030*"aika"</b>
0.017*"tulla"	0.018*"sitten"	<b>0.018*"masentua"</b>	<b>0.028*"muutama"</b>
0.016*"yhtä"	0.018*"väli"	<b>0.017*"diagnoosi"</b>	<b>0.024*"kuukausi"</b>
<b>0.015*"helpottaa"</b>	0.018*"kauan"	<b>0.016*"saada"</b>	0.024*"kaksi"
AUKTORITEETTINEUVO #21	AUKTORITEETTINEUVO #49	FRAASI #24	FRAASI #39
<b>0.157*"lääkäri"</b>	<b>0.088*"apu"</b>	0.240*"mikään"	0.129*"kerta"
<b>0.055*"löytyä"</b>	<b>0.050*"saada"</b>	0.192*"hyvä"	0.088*"sä"
<b>0.048*"sairaus"</b>	<b>0.048*"vastata"</b>	<b>0.045*"kiitos"</b>	<b>0.067*"kokeilla"</b>
<b>0.039*"hoitaa"</b>	<b>0.035*"hakea"</b>	<b>0.030*"vastaus"</b>	0.039*"tarkoittaa"
<b>0.035*"todeta"</b>	0.030*"toivoa"	0.023*"yhtään"	0.033*"ensimmäinen"
0.023*"käydä"	0.027*"voida"	0.021*"nimi"	0.032*"sopia"
0.020*"voida"	0.026*"jatkaa"	0.018*"saada"	<b>0.019*"tsemppi"</b>
0.019*"vika"	0.024*"raskas"	0.018*"vaihtoehto"	0.016*"kaikki"
0.013*"päin"	<b>0.019*"apteekki"</b>	0.018*"kaikki"	0.015*"maha"
0.013*"yksityinen"	0.017*"koti"	0.015*"muu"	<b>0.013*"kannattaa"</b>



SAIRAUDET #2	SAIRAUDET #4	SAIRAUDET #6	SAIRAUDET #19
0.088*"maailma"	0.081*"%"	0.053*"kysymys"	<b>0.062*"pilleri"</b>
0.068*"liikaa"	0.035*"vaarallinen"	0.044*"niisi"	0.058*"puoli"
<b>0.047*"e-pilleri"</b>	0.027*"100"	0.041*"vaihtaa"	0.050*"elää"
<b>0.029*"verikoe"</b>	<b>0.024*"päänsärky"</b>	<b>0.032*"sairastua"</b>	0.039*"luulla"
0.026*"vaikkaei"	0.024*"30"	<b>0.026*"aivo"</b>	0.029*"laskea"
0.024*"kylmä"	0.024*"muttaei"	0.023*"ikävä"	<b>0.025*"menkat"</b>
0.019*"aurinko"	<b>0.023*"ylipainoinen"</b>	<b>0.019*"riski"</b>	0.020*"tulla"
0.017*"käyttää"	0.022*"tuska"	<b>0.019*"välttää"</b>	0.017*"oppia"
0.014*"selkeä"	<b>0.022*"alkoholisti"</b>	<b>0.019*"lisääntyä"</b>	<b>0.016*"vuoto"</b>
<b>0.011*"homeopaatti"</b>	0.020*"todennäköisesti"	<b>0.018*"seuraus"</b>	0.016*"arvo"
SAIRAUDET #22	SAIRAUDET #28	SAIRAUDET #36	SAIRAUDET #42
0.067*"suuri"	0.065*"lopettaa"	0.532*"ottaa"	<b>0.065*"käsi"</b>
0.052*"osa"	<b>0.052*"pää"</b>	0.075*"kiinni"	<b>0.042*"jalka"</b>
0.042*"liian"	<b>0.034*"rinta"</b>	<b>0.026*"homeopatia"</b>	0.042*"paremmin"
0.031*"kuolla"	0.031*"alkaa"	<b>0.024*"tuntemus"</b>	0.040*"meno"
0.031*"kovin"	<b>0.027*"loppua"</b>	0.023*"vastuu"	<b>0.026*"vaivata"</b>
<b>0.029*"nukkua"</b>	<b>0.024*"voima"</b>	0.014*"kelvata"	0.024*"päällä"
<b>0.027*"yö"</b>	0.024*"sitten"	0.009*"kiltti"	<b>0.022*"laihutus"</b>
<b>0.026*"uni"</b>	0.024*"kasvaa"	<b>0.009*"hillitä"</b>	0.020*"selvästi"
0.023*"lihava"	0.017*"kävellä"	<b>0.008*"oireeton"</b>	0.019*"koira"
0.019*"samanlainen"	0.017*"nostaa"	0.008*"jep"	0.016*"erityisesti"
SAIRAUDET #48	NARRATIIVI #1	NARRATIIVI #7	NARRATIIVI #8
0.062*"enemmän"	<b>0.077*"toimia"</b>	0.083*"silmä"	<b>0.122*"lääke"</b>
0.048*"vaikea"	<b>0.054*"aine"</b>	0.072*"pysyä"	0.050*"osata"
<b>0.038*"hammas"</b>	<b>0.040*"elimistö"</b>	0.067*"tää"	0.032*"jokainen"
0.037*"hetki"	0.029*"kuukautinen"	0.048*"ns"	<b>0.023*"selvitä"</b>
0.028*"viesti"	0.018*"millainen"	<b>0.033*"epäillä"</b>	0.023*"tee"
<b>0.024*"suu"</b>	0.017*"selittää"	0.031*"ihana"	0.021*"saada"
0.020*"vähemmän"	0.016*"voida"	0.028*"keksiä"	<b>0.021*"määrätä"</b>
0.018*"toisaalta"	0.016*"hyvin"	0.027*"outo"	<b>0.019*"neuvo"</b>
0.015*"tarpeeksi"	<b>0.015*"tehota"</b>	0.025*"henki"	0.019*"pyytää"
0.013*"usko"	0.014*"väärin"	0.023*"kevyt"	0.019*"kyseinen"
NARRATIIVI #11	NARRATIIVI #15	NARRATIIVI #27	NARRATIIVI #29
<b>0.103*"kertoa"</b>	<b>0.082*"hoito"</b>	<b>0.064*"oire"</b>	<b>0.080*"pois"</b>
<b>0.088*"kukaan"</b>	<b>0.059*"tutkimus"</b>	<b>0.054*"aiheuttaa"</b>	<b>0.063*"lähteä"</b>
<b>0.070*"kokemus"</b>	<b>0.042*"vaikutus"</b>	0.033*"lisäksi"	<b>0.038*"jättää"</b>
<b>0.056*"tieto"</b>	0.030*"poistaa"	0.025*"voida"	0.032*"aamu"
<b>0.050*"normaali"</b>	<b>0.028*"vaatia"</b>	<b>0.025*"lääkitys"</b>	0.031*"ilta"
0.027*"tällainen"	0.022*"ollenkaan"	0.022*"myös"	0.024*"tänään"
0.026*"voida"	0.019*"psykiatri"	0.018*"mm"	0.024*"lisää"
0.022*"muu"	<b>0.017*"parantaa"</b>	0.017*"ilmeisesti"	<b>0.023*"viedä"</b>
0.019*"uskaltaa"	<b>0.014*"teho"</b>	0.017*"vatsa"	<b>0.022*"sairaala"</b>
0.018*"muutos"	0.013*"järki"	0.015*"keho"	0.021*"saada"

NARRATIIVI #33	NARRATIIVI #43	NARRATIIVI #45	NARRATIIVI #47
<b>0.070**huomata</b>	<b>0.071**huono</b>	<b>0.130**auttaa</b>	<b>0.098**ongelma</b>
<b>0.038**kipu</b>	<b>0.066**tulos</b>	<b>0.061**ymmärtää</b>	<b>0.048**löytää</b>
0.030**tosiaan	0.060**jne	0.049**kyse	0.030**jokin
0.029**riittää	<b>0.043**terve</b>	0.043**kuulua	<b>0.028**toivottavasti</b>
0.029**pakko	0.033**voi	<b>0.038**potilas</b>	0.026**keskustelu
0.027**muuttua	0.024**alle	0.030**kuitenkaan	0.025**vaihe
0.025**oikeasti	0.022**tällöin	0.026**voida	0.024**tuttu
0.024**tavallinen	0.022**asua	0.019**ikinä	0.022**väittää
<b>0.022**parata</b>	0.016**silla	<b>0.015**hyväksyä</b>	0.019**kuulostaa
0.022**tajuta	<b>0.013**mieliala</b>	<b>0.013**itkeä</b>	0.018**kauppa
META #34	META #10	META #41	N/A #35
0.084**mies	0.239**..	0.090**syy	0.039**the
0.065**nainen	0.062**lukea	0.079**lapsi	0.035**just
0.050**muistaa	0.049**kirjoittaa	0.043**vanha	0.033**of
0.038**palsta	0.042**kysyä	0.043**äiti	0.026**tuottaa
0.035**tapahtua	0.037**maksaa	0.023**isä	0.025**and
0.030**nuori	0.027**toi	0.021**sairas	0.022**e
0.029**ikä	0.022**turha	0.021**poika	0.021**in
0.024**ym	0.019**ainoa	0.020**samoin	0.019**i
0.016**tyttö	0.019**aihe	0.019**kotona	0.018**to
0.015**tietenkin	0.017**kirjoitus	0.019**eräs	0.018**minkäänlainen
topic #12	topic #14	topic #18	topic #20
0.128**esim	0.084**aivan	0.066**uskoa	0.066**heti
0.027**koulu	0.061**n	0.046**eri	0.046**lisätä
0.027**tms	0.043**tuoda	0.040**=	0.043**alku
0.021**voida	0.040**kiva	0.028**yksin	0.031**melkein
<b>0.020**selkä</b>	0.038**sattua	0.025**nauttia	0.030**tarve
0.018**kuvitella	0.031**varma	0.024**takaisin	<b>0.028**seurata</b>
0.018**ajaa	0.031**sieltä	0.020**keino	<b>0.025**sairastaa</b>
0.016**auto	0.029**sivu	0.018**kunnolla	0.024**kuva
0.013**riippuvuus	0.017**energia	<b>0.017**nälkä</b>	0.020**merkki
0.012**musta	0.016**hirveä	0.016**saada	0.019**nimittäin
topic #23	topic #25	topic #30	topic #31
0.165**tietää	0.199**mä	0.142**sama	0.060**puhua
0.119**haluta	0.112**sanoa	0.054**tapa	0.054**juoda
0.049**no	0.051**tosi	0.049**täällä	0.054**juttu
0.025**onnistua	0.029**ku	0.047**kuulla	0.035**ystävä
0.024**yhteys	0.026**mut	0.032**täytyä	0.035**kaveri
<b>0.021**laihduttaa</b>	0.022**sit	0.031**hei	<b>0.034**liikunta</b>
0.017**voida	0.018**voida	0.025**ensin	0.024**vesi
0.014**vain	0.018**tulla	0.020**mukava	<b>0.022**syöpä</b>
0.014**menettää	0.017**kyllä	0.019**homma	0.019**harrastaa
0.013**mikäli	<b>0.016**leikkaus</b>	0.019**muu	0.015**lasi

topic #37	topic #38
0.211*"tehdä"	0.100*"päästä"
0.049*"työ"	0.059*"johtua"
0.033*"näyttää"	0.035*"taitaa"
0.027*"mieli"	0.033*"ero"
<b>0.022*"tutkia"</b>	0.025*"mainita"
0.019*"voida"	0.023*"voida"
0.016*"kesä"	0.019*"nopeasti"
0.015*"jokin"	<b>0.018*"finni"</b>
0.015*"pitää"	<b>0.017*"sivuvaikutus"</b>
0.011*"asia"	0.015*"korkea"